PREPRO INSULIN-LIKE GROWTH FACTORS I AND II

This application is a continuation of application Ser. No. 630,557, filed 13 Jul. 1984, now abandoned.

BACKGROUND OF THE INVENTION

1. Field of the Invention

It is suspected that somatic growth which follows the administration of growth hormones in vivo is mediated through a family of mitogenic, insulin-like peptides whose serum concentrations are growth hormone dependent. These polypeptides include somatomedin-C, somatomedin-A, and insulin-like growth factors I and II (IGF-I and IGF-II). IGF-I and IGF-II are single chain serum proteins of 70 and 67 amino acids, respectively, and there is evidence that they are identical to somatomedin-C and somatomedin-A. Although IGF-I and IGF-II can be isolated from human serum, such separation at best provides only limited quantities of the growth factors. It would thus be of great scientific and clinical interest to be able to produce relatively large quantities of the growth factors by recombinant DNA techniques. In order to do so, it is necessary to have DNA sequences which encode for IGF-I and IGF-II. In particular, it would be desirable to derive such DNA sequences from their natural source, i.e., human genetic information (RNA or DNA).

2. Description of the Prior Art

The amino acid sequences for human insulin-like growth factors I and II were first determined by Rinderknecht and Humbel (1978) J. Biol. Chem. 253:2769–2776 and Rinderknecht and Humbel (1978) FEBS. Lett. 89:283–286, respectively. The chemical synthesis of biologically active IGF-I has been reported. Li et al. (1983) Proc. Natl. Acad. Sci. USA 80:2216–2220. See also copending application Ser. No. 487,950, filed Apr. 25, 1983, which discloses the expression of synthetic genes for IGF-I and IGF-II in yeast.

SUMMARY OF THE INVENTION

Nucleotide sequences including both DNA and RNA are provided which code for human insulin-like growth factors (IGF) I and II and their corresponding polypeptide precursors. The DNA sequences may be used for production of the IGF and precursor polypeptides and biologically-active portions thereof in microorganisms or cell culture, while both the DNA and RNA sequences are useful as labelled probes in detecting the presence of the growth factor genes and/or mRNA sequences in a natural source. The nucleotide sequences of the present invention are derived from genetic information isolated from human cells, typically liver cells. In the exemplary embodiment, a cDNA library derived from human liver cells is screened with radiolabelled hybridization probes encoding a short nucleotide sequence common to both IGF-I and IGF-II. In this way, DNA sequences encoding for both preproIGF-I and preproIGF-II were detected and isolated.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 sets forth the nucleotide sequence derived from plasmid phigf1 encoding human preproIGF-I. The predicted amino acid sequence of the prepro protein is provided, and the first amin acid of the mature protein is designated as number 1. The regi n corre-

sponding to mature IGF-I is boxed, and pairs of basic amino acids are underlined.

FIG. 2 sets forth the nucleotide sequence derived from plasmid phigf2 encoding human preproIGF-II. The predicted amino acid sequence of the prepro protein is numbered with the first amino acid designated as number −24. The region corresponding to mature IGF-II is boxed, and pairs of basic amino acids are underlined.

FIG. 3 is a schematic representation of the structure of preproIGF-II. The proteolytic processing site of proIGF-II is indicated by an arrow; K and R denote lysine and arginine, respectively.

FIG. 4 is a schematic representation of the structure of preproIGF-I. The proteolytic processing site of proIGF-I is indicated by an arrow; K and R denote lysine and arginine, respectively.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

According to the subject invention, DNA and RNA sequences encoding for human IGF-I and IGF-II prepro polypeptides, or portions thereof, are provided. Such nucleotide sequences are useful for a variety of purposes. Both DNA and RNA sequences including at least 12 bases, more usually at least 18 bases, and frequently having 50 bases or more, can be used as hybridization probes for detecting complementary sequences in genomic DNA or in messenger RNA. Such probes may be used for detecting mutations and/or deletions in humans suspected of suffering from growth deficiencies. Longer DNA sequences may be used for expressing the precursor and/or mature proteins incorporating IGF, or any fragments or analogs thereof. Production of the precursor polypeptides will often be desirable since the precursor will be amenable to post-translational processing in appropriate hosts. The DNA sequences may also be used for the production of mRNA for any of the above purposes.

Both IGF-I and IGF-II are initially translated as "prepro" polypeptides including an amino-termino signal peptide and a carboxy-terminal peptide, referred to as the E domain. The signal peptide directs secretion of the prepro polypeptide across intracellular membranes and is cleaved during such secretion to form the "pro" polypeptide. Mature IGF-I and IGF-II are formed by subsequent proteolysis of the carboxy-terminal E domain from the pro polypeptide.

The nucleotide sequences of the present invention will be derived from human cells, typically by screening a human cDNA or genomic DNA library with hybridization probes capable of detecting a nucleotide sequence predicted from the known amino acid sequences of IGF-I and II. While suitable genomic libraries may be derived from human cells of any origin, it is preferred to utilize cDNA libraries from cells which are known to express the insulin-like growth factors, such as human liver cells and human fetal cells. Suitable hybridization probes may be synthesized by well known techniques and should employ degenerate coding to provide for all possible codons corresponding to each amino acid. In the exemplary embodiment, a human liver cDNA library developed by Woods et al. ((1982) Proc. Natl., Acad. Sci. USA 79:5661–5665) was screened with a 23 base ligonucleotide probe based on an 8 amino acid sequence common to both IGF-I and IGF-II.

The DNA sequences of interest in the present invention may be single or double stranded and will include at least about 12 bases, preferably 18 bases or more, for single stranded oligonucleotides useful as hybridization probes. Double stranded fragments used for expression of polypeptides will usually be longer, typically being at least 18 base pairs corresponding to a sequence of 6 amino acids, more typically being at least the length of the coding region for the mature polypeptide, or a physiologically active fragment thereof. The DNA sequences may extend the entire length of the coding region for the prepro polypeptide, and may include untranslated and/or untranscribed flanking regions on either side of such coding region and/or intervening sequences.

Once the IGF DNA of interest has been isolated from the human cellular source, it will usually be cloned and expanded to provide sufficient amounts of the DNA for the intended use. Once sufficient amounts of the DNA have been obtained, the DNA sequence may be modified in a number of ways. For example, DNA sequences used as hybridization probes will be cleaved to a desired length using restriction enzymes, denatured to single-stranded form, and labelled, typically with a radiolabel, to allow detection. For expression of the mature IGF polypeptides, it may be desirable to excise the coding regions for the mature polypeptide and insert such coding regions into a suitable expression vector. In this way, the mature polypeptide may be expressed in hosts which are incapable of processing the prepro or pro polypeptide. Alternatively, in suitable hosts it may be desirable to employ the coding region for the entire prepro polypeptide either with or without associated flanking or intervening sequences.

The DNA sequences of the present invention may be replicated and expressed in a wide variety of hosts, including prokaryotes, eukaryotes, and mammalian cell culture. The cDNA sequences may be introduced into the host by conventional techniques, usually employing an extrachromosomal element capable of stable replication within the host. Alternatively, the DNA may be introduced directly into the genomic DNA using, e.g., co-transformation as described by Wigler et al. (1979) Cell 16:777–785. Hosts of particular interest include unicellular microorganisms, such as *E. coli, S. cerevisiae,* and *B. subtilis.*

A wide variety of suitable extrachromosomal elements exist for the cloning and expression of the IGF DNA sequences of the present invention. The cloning vectors will be selected to include a replication system suitable for the intended host. Suitable expression vectors for mammalian cells are well known in the art and include those having replication systems derived from viral genomes or portions thereof, e.g., SV-40, retroviruses, and the like. Replication systems for *E. coli* include those derived from various plasmids, such as R6-5, ColE1, RSF, and the like. Particularly convenient is plasmid pBR322 which includes a replication system derived from pMV1. Vectors suitable for yeast include those having a replication system derived from the 2 μm plasmid, autonomously replicating sequences (ars), and the like. Frequently, it will be desirable to have replication systems for both *E. coli* and a higher organism, e.g., yeast, present on the sam extrachromosomal element. Such vect rs, referred to as shuttle vectors, allow for cloning and amplification of the IGF gene in bacteria, while expression may be achieved in the higher organism with appropriate RNA or post-transla-

tional processing, e.g., cleavage of the pro polypeptide at the appropriate site to yield the mature polypeptide, polyadenylation, splicing, and the like.

In addition to the replication system, suitable extrachromosomal elements will usually include at least one marker f r each intended h st cell which allows for selection or selective pressure to maintain the extrachromosomal element containing the IGF DNA sequence. Convenient markers include biocidal resistance, e.g., antibiotics, heavy metals and toxins; complementation in an auxotrophic host, and the like.

The following examples are offered by way of illustration and not by way of limitation.

EXPERIMENTAL

Methods

Transformants (approximately 9000) from the adult human liver cDNA library of Woods et al. (1982) Proc. Natl. Acad. Sci. USA 79:5661–5655 were grown in 96-well microtiter dishes. The cDNA library was constructed by inserting ds cDNA prepared from mRNA from adult human liver cells into the PstI restriction site of pKT218. Plasmid pKT218 is a pBR322 derivative described by Talmadge et al. (1980) Proc. Natl. Acad. Sci. USA 77:3369–3373. Colonies of transformed *E. coli* were transferred to Whatman 541 paper, grown, amplified with chloramphenicol, and lysed as described by Gergen et al. (1979) Nucleic Acids Res. 7:2115–2136. Colonies containing IGF sequences were identified by hybridization with a 256-fold degenerate 23 base oligonucleotide which had been labelled with [γ-^{32}P]-ATP and T4 polynucleotide kinase. The oligonucleotide was synthesized manually (Urdea et al. (1983) Proc. Natl. Acad. Sci. USA 80:7461–7465) and purified by electrophoresis in a 8M urea, 20% polyacrylamide gel. The filters were hybridized in 5XSSC (SSC is 0.15M NaCl, 0.015M sodium citrate), 50 mM sodium phosphate, pH 7.0, 0.2% sodium dodecyl sulfate (SDS), 2X Denhardt's (Denhardt (1966) Biochem. Biophys. Res. Commun. 23:641–646), 200 μg/ml, sonicated and denatured salmon testes DNA, and 10^6 cpm/ml of 32P-labelled oligonucleotide at 30° C. After 16 hr, the filters were washed in 5XSSC and 0.1% SDS at 42° C. for one hour. Hybridizing colonies were identified by autoradiography. The inserted DNA fragments in the plasmids from the hybridizing colonies were sequenced. The sequence of the fragments carrying IGF-I or IGF-II DNA were determined on both strands and across all restriction sites used to initiate sequence determinations by the procedures of Maxam and Gilbert supra. and Sanger et al. (9180) J. Mol. Biol. 143:161–178.

Results

The nucleotide sequence of the hybridization probe was based on an eight amino acid sequence common to the sequences of IGF-I (amino acids 46–53) and IGF-II (amino acids 45–52), as reported by Rinderknecht and Humbel (1978) J. Biol. Chem. 253:2769–2776 and FEBS Lett. 89:283–286. The sequence was as follows.

	Glu	Cys	Cys	Phe	Arg	Ser	Cys	Asp	
5'-GAG	TGT	TGT	TTT	CGA	AGT	TGT	GA-3'		
	A	C	C	C	T	C	C		
					G				
					C				

Eight of the approximately 9000 colonies hybridized with this probe, and analyses of the inserted PstI fragments derived fr m the cloned plasmids revealed that

only four of the eight fragments were of different sizes. The nucleotide sequences of each of the four distinct fragments were determined, and the predicted amino acid sequences corresponding to each of the three reading frames compared with the known sequences of IGF-I and IGF-II. The fragments in two of the plasmids (designated phigf 1 and phigf 2) were found to encode IGF-I and IGF-II, respectively. The proteins encoded by the other fragments were not related to IGF.

The plasmid phigf 1 contained a PstI insert of approximately 660 base pairs (bp) which extended from the second nucleotide of the codon for amino acid −15 of the putative signal peptide of the preproIGF-I to the poly A tract and included about 245 bp of 3' untranslated region and a variant polyadenylation signal AATAAT (residues 595–600). The nucleotide sequence and corresponding amino acid sequence are set forth in FIG. 1. The number of the nucleotide at the end of each line is indicated; the region corresponding to mature IGF-I is boxed and pairs of basic amino acids are underlined.

The nucleotide sequence of human preproIGF-II mRNA was deduced from the sequence of the inserted fragment in plasmid phigf 2. Referring to FIG. 2, the predicted amino acid sequence of preproIGF-II is numbered with the first amino acid of preproIGF-II designated as number −24. The region corresponding to mature IGF-II is boxed and pairs of basic amino acids are underlined. An 89 amino acid carboxy-terminal region comprises residues 68–156. The number of the nucleotide at the end of each line is indicated. The B-domain (FIG. 3) of IGF-II comprises residues 1–32, the C-domain comprises residues 33–40, the A-domain comprises residues 41–61, the D-domain comprises residues 62–67, and the carboxyl-terminal E-domain comprises residues 68–156. In comparison, the B-domain of IGF-I (FIG. 4) comprises residues 1–29, the C-domain comprises residues 30–41, the A-domain comprises residues 42–62, the D-domain comprises residues 63–70, and the carboxy-terminal E-domain comprises residues 71–105.

Translation of the IGF-II mRNA from the initial Met (nucleotides 251–253, FIG. 2) predicts an 180 amino acid protein in which the 67 amino acid sequence of IGF-II begins 25 residues from the start. Thus, including the opal termination codon, the coding region is 543 bases. The 5'-untranslated region of the mRNA is at least 250 bases, and the 3'-untranslated region is greater than 253 bases. The cDNA clone phigf 2 lacks a poly A tract and polyadenylation signal.

No other clones encoding preproIGF-II mRNA were revealed when the insert in phigf 2 was used as a probe to rescreen the original 9000 colonies and 6000 additional colonies. A similar experiment using the insert of phigf1 as a probe to screen the same 15,000 colonies revealed, besides phigf1, only a second, identical clone, previously detected in the original screen, and phigf2 which cross-hybridized weakly. Attempts to determine the sizes of human preproIGF-I and preproIGF-II mRNAs by hybridization of the inserts to a northern blot (Thomas (1980) Proc. Natl. Acad. Sci. USA 77:5201–5205) of human adult liver poly A+ RNA were inconclusive, presumably because f the low abundance of these mRNAs (<1/10,000 molecules) in this tissue.

Both IGF-I and IGF-II are secreted proteins, and the 24 residue amino-terminal extension of the latter ap-

pears to be the signal peptide. Analysis of the hydr philicity of preproIGF-II (as described by Hopp and Woods (1981) Proc. Natl. Acad. Sci. USA 78:3824–3828) indicates that the putative signal peptide has a hydrophobic core of 14 residues (amino acids −15 to −2) and a profile similar to other signal peptides. It is concluded that the homologous amino-terminal extension of preproIGF-I also represents a signal peptide of at least 15 amino acids. Interestingly, about 25% of the purified human IGF-II molecules lack Ala 1 (Rinderknecht and Humbel (1978) FEBS Lett. 89:283–286) suggesting that cleavage of the Ala(−1)-Ala(1) peptide bond by the peptidase is preferred but that the Ala(1-)-Tyr(2) bond is also cleaved.

The ≧15 and 24 residue amino-terminal extensions are cleaved from preproIGF-I and preproIGF-II, respectively, to produce proIGF-I and proIGF-II. ProIGF-II includes the 89 amino acid carboxyl-terminal extension referred to as the E-domain, and proteolytic processing at Arg 68 is required to produce mature IGF-II. Similarly, proIGF-I contains an E domain but of only 35 amino acids with requisite proteolytic processing to produce mature IGF-I at Arg 71. This carboxyl-terminal extension also has a potential N-linked glycosylation site (residues 92–94:Asn-Ala-Ser), which is absent in the IGF-II precursor. Although proteolytic processing at single basic residues has been reported in the generation of other proteins, including epidermal growth factor (Scott et al. (1982) Science 221:236–240) and growth hormone releasing factor (Gubler et al. (1983) Proc. Natl. Acad. Sci. USA 80:4311–4314; Mayo et al. (1983) Nature 306:86–88), processing occurs more often at pairs of basic amino acids. Only two such sites occur in proIGF-I (indicated by underlining in FIG. 1), both of which are within the mature polypeptide (residues 36–37 and 55–56), and thus remain uncleaved. There are five paired basic amino acids in proIGF-II (underlined in FIG. 2) including one site within mature IGF-II (residues 37–38) that is not cleaved. It is unknown if proteolysis occurs at any of the other pairs of basic residues (proIGF-II) or single basic amino acids within either of the E-domains.

In accordance with the subject invention, polynucleotide sequences are provided which encode insulin-like growth factors I and II. The polynucleotide sequences are derived from human genetic information (either DNA or RNA), typically by screening a cDNA library with an appropriate hybridization probe, and are useful for expression of the prepro polypeptide as well as the mature polypeptide. Additionally, the cloned polynucleotides themselves may be labelled and used as hybridization probes for a variety of purposes, such as genetic screening.

Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be obvious that certain changes and modifications may be practiced within the scope of the appended claims.